CSE574 Introduction to Machine Learning

Adversarial Attack: An Overview

Jue Guo

University at Buffalo

March 31, 2024

Outline

What are adversarial attacks?

- The surprising findings by Szegedy (2013) and Goodfellow (2014)
- Example of attacks
- Physical Attacks

Basic Terminologies

Multi-class Problem

Why?

- Robustness = easiness to fail when input is perturbed. Perturbation can be in any kind. Robustness machine learning is a very rich topic.
- We will look at something very narrow, called **adversarial robusness**, also known as robustness against **attacks**.
- Adversairal attack is a very **hot** topic, as of today. We should not over-emphasize its importance. There are many other important problems.

Adversarial Attack Example: FGSM

- It is not difficult to fool a classifier
- The perturbation could be perceptually not noticeable

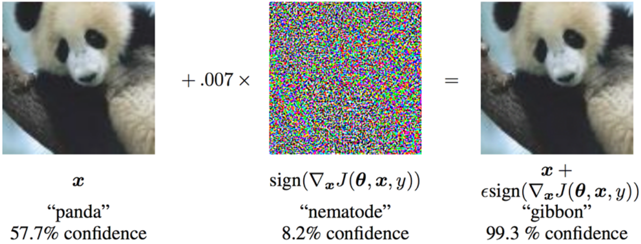

$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Figure: Goodfellow et al. "Explaining and Harnessing Adversarial Examples", https://arxiv.org/pdf/1412.6572.pdf

Adversarial Attack Example: Szegedy's 2013 Paper

- This paper actually appears one year before Goodfellow's 2014 paper.

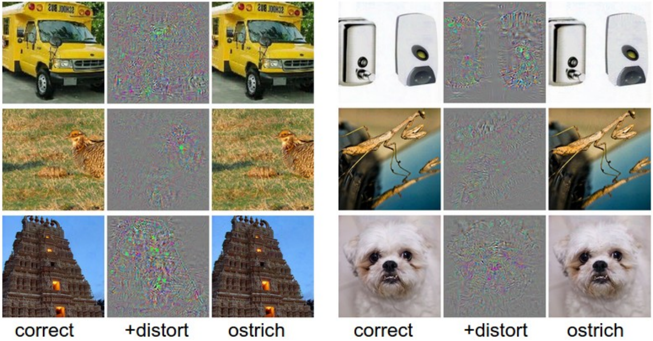


Figure: Szegedy et al. Intriguing properties of neural networks
https://arxiv.org/abs/1312.6199

Adversarial Attack: Targeted Attack

- Targeted Attack

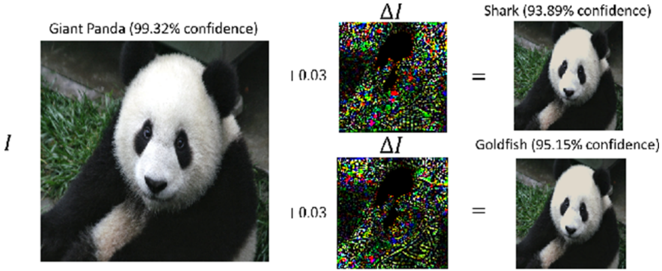

Giant Panda (99.32% confidence)
I
+ 0.03
ΔI
=
Shark (93.89% confidence)
ΔI
+ 0.03
=
Goldfish (95.15% confidence)

Figure: Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics, https://arxiv.org/abs/1612.07767

Adversarial Attack Example: One Pixel

- One-pixel Attack

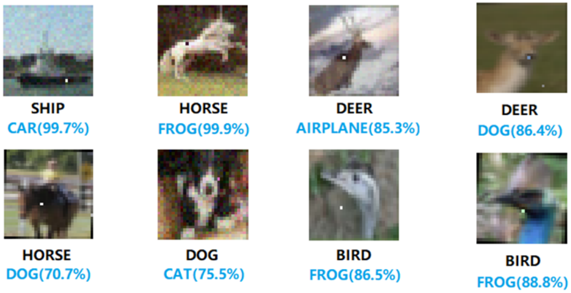
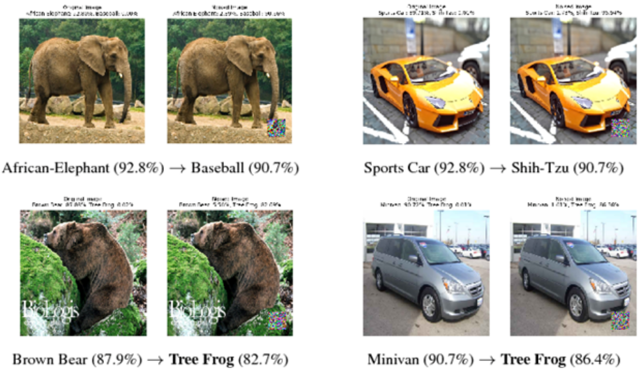


Figure: One pixel attack for fooling deep neural networks
https://arxiv.org/abs/1710.08864

Adversarial Attack Example: Patch

- Adding a patch



African-Elephant (92.8%) → Baseball (90.7%)

Sports Car (92.8%) → Shih-Tzu (90.7%)

Brown Bear (87.9%) → **Tree Frog** (82.7%)

Minivan (90.7%) → **Tree Frog** (86.4%)

Figure: LaVAN: Localized and Visible Adversarial Noise, https://arxiv.org/abs/1801.02608

Adversarial Attack Example: Stop Sign

- The Michigan / Berkely Stop Sign

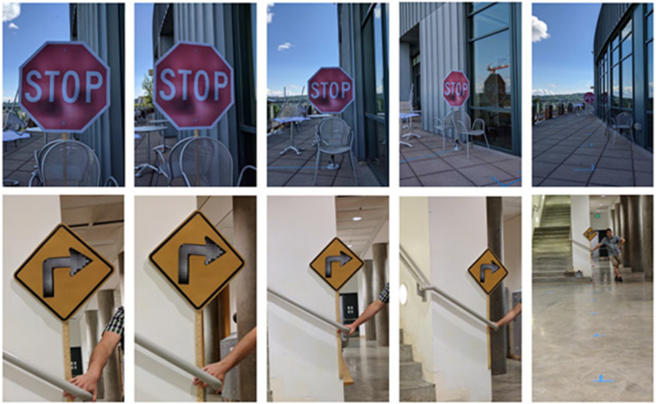


Figure: Robust Physical-World Attacks on Deep Learning Models
https://arxiv.org/abs/1707.08945

Adversarial Attack Example: Turtle

- The MIT 3D Turtle

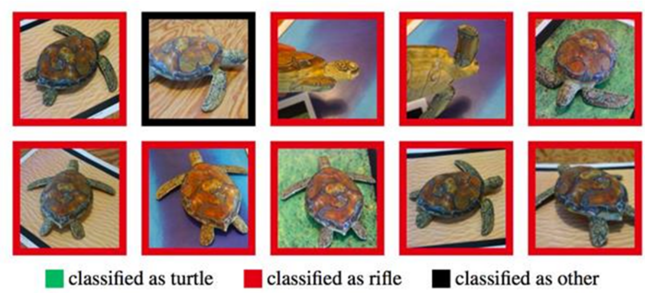


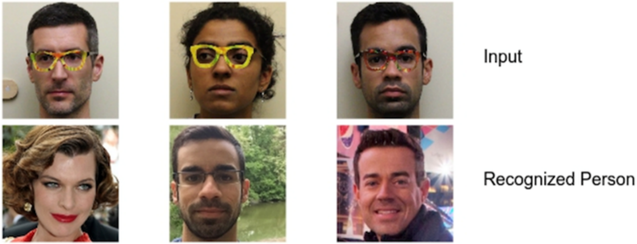
Figure: Synthesizing Robust Adversarial Examples

https://arxiv.org/pdf/1707.07397.pdf

https://www.youtube.com/watch?v=YXy6oX1iNoA

Adversarial Attack Example: Glass

- CMU Glass



Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016, October).
Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition.
In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1528-1540). ACM.

Figure: Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition https://www.cs.cmu.edu/ ~sbhagava/papers/face-rec-ccs16.pdf https://www.archive.ece.cmu.edu/ ~lbauer/proj/advml.php

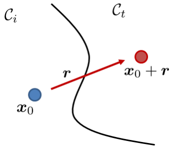
Definition: Additive Adversarial Attack

Additive Adversarial Attack

Let $x_0 \in \mathbb{R}^d$ be a data point belong to class $\mathscr{C}_i$. Define a target class $\mathscr{C}_t$ An additive adversarial attack is an addition of a perturbation $r \in \mathbb{R}^d$ such that the perturbed data

$$x = x_0 + r$$

is misclassified as $\mathscr{C}_t$.



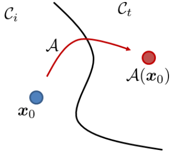
Definition: General Adversarial Attack

General Adversarial Attack

Let $x_0 \in \mathbb{R}^d$ be a data point belong to class $\mathscr{C}_i$. Define a target class $\mathscr{C}_t$ An adversarial attack is a mapping $\mathscr{A} : \mathbb{R}^d \to \mathbb{R}^d$ such that the perturbed data

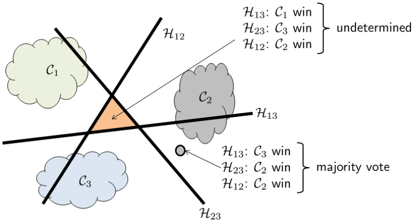
$$x = \mathscr{A}(x_0)$$

is misclassified as $\mathscr{C}_t$.



Multi-class Problem

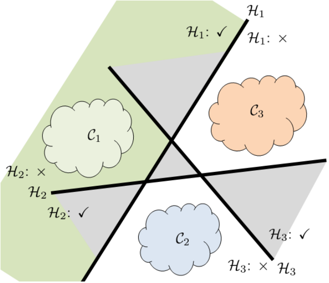
Approach 1: One-on-One


$\mathcal{C}_1$
$\mathcal{C}_2$
$\mathcal{C}_3$
$\mathcal{H}_{12}$
$\mathcal{H}_{13}$
$\mathcal{H}_{23}$
$\mathcal{H}_{13}$: $\mathcal{C}_1$ win
$\mathcal{H}_{23}$: $\mathcal{C}_3$ win
$\mathcal{H}_{12}$: $\mathcal{C}_2$ win
undetermined
$\mathcal{H}_{13}$: $\mathcal{C}_3$ win
$\mathcal{H}_{23}$: $\mathcal{C}_2$ win
$\mathcal{H}_{12}$: $\mathcal{C}_2$ win
majority vote

- Class i vs. Class j
- Give me a point, check which class has more votes
- There is an undetermined region

The Multi-Class Problem

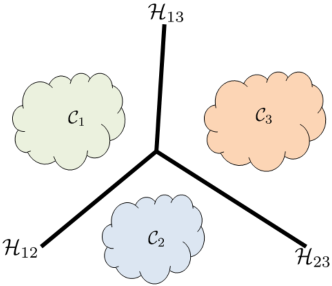
Approach 2: One-on-All



- Class i not Class i
- Give me a point, check which class has no conflict
- There are undetermined regions

The Multi-Class Problem

Approach 3: Linear Machine



- Every point in the space gets assigned a class.
- You give me x, I compute $g_1(x), g_2(x), \ldots, g_K(x)$
- If $g_i(x) \geq g_j(x)$ for all $j \neq i$, then x belongs to class i